Nucleic Acids Research, 2006, Vol. 34, Web Server issue ***W133–W136***
doi:10.1093/nar/gkl251

MAVL/StickWRLD: analyzing structural constraints using interpositional dependencies in biomolecular sequence alignments

Hatice Gulcin Ozer and William C. Ray[1,*]

Biophysics Graduate Program and [1]Children's Research Institute and The Department of Pediatrics, The Ohio State University, 700 Children's Drive. Columbus, OH 43205, USA

Received February 14, 2006; Revised March 1, 2006; Accepted March 30, 2006

ABSTRACT

The increasing availability of structurally aligned protein families has made it possible to use statistical methods to discover regions of interpositional dependencies of residue identity. Such dependencies amongst residues often have structural or functional implications, and their discovery can supply valuable constraints that assist in the refinement of measured, or predicted molecular structure assignments. Multiple Alignment Variation Linker (MAVL) and StickWRLD [W. Ray (2004) *Nucleic Acids Res.*, 32, W59–W63] were developed to analyze and visualize nucleic acid and protein alignments, to discover and illuminate position/location relationships to the user. The original system analyzed users' data from a web-form submission and presented the user with a static VRML diagram describing their data. We are pleased to report that MAVL/StickWRLD has been completely redesigned and rewritten. MAVL/StickWRLD now functions as a platform-independent Java applet, with real-time dynamic controls that enable much more intuitive exploration and interaction with the data. The system has also been upgraded to enable visualization of a range of aggregate residue properties, and an extensive database of pre-computed StickWRLD diagrams based on PFAM families is now available directly from the interface. The Java StickWRLD applet is available via the WWW at http://www.microbial-pathogenesis.org/stickwrld/.

INTRODUCTION

The determination of a biomolecule's complete three-dimensional structure provides invaluable information about the molecule's function, and how that function might be modulated. This in turn leads to insights across the biological spectrum. Unfortunately, determining protein or nucleic-acid structures from the molecules themselves is difficult owing to physical impediments. Likewise, accurate computational prediction of such structures from physical principles remains computationally intractable owing to the overwhelming number of degrees of freedom in the system. Techniques that can determine constraints on the potential molecular conformations can dramatically improve the chances of determining structures with both measured and predictive methods.

One possible method for predicting physical proximity is to examine large families of sequences that have presumably similar structures, and to catalog positional pairs that show correlation between the identities of the residues occupying them. This calculation produces a four-dimensional matrix of positions, identities, related positions and strengths of relationship. For a small RNA molecule of 80 nt, this 4D matrix contains 160 000 values, while a small protein of 300 amino acids produces a cross-correlation matrix containing 39 690 000 values.

Since there is no single value or distinctly definable pattern amongst these values that signals a structural relationship, an abstraction that allows the expert researcher to visualize, interpret and query the matrix is clearly necessary (http://www.siggraph.org/s2005/main.php?f=conference&p=posters).

MAVL/StickWRLD was developed to visualize interpositional dependencies within nucleic acid and amino acid sequence alignments (1,2). In this representation, a positional weight matrix representing the alignment is wrapped around a cylinder by placing spheres for each identity and position. The diameters of the spheres are proportional to the percentage frequency of the corresponding identity. Statistically significant pairwise correlations are depicted as sticks between the related pairs of spheres. The diameter of the sticks is based on the over-representation or under-representation of sequences sharing these identities as compared with a weight-matrix-based expectation for the sequences as a family. This visualization method allows the researcher to detect patterns

*To whom correspondence should be addressed. Tel: +1 614 355 3522; Fax: +1 614 722 2818; Email: ray.29@osu.edu

of pairwise positional dependencies easily, and comment on possible structural implications of these observations.

In this paper, we present a complete rewrite of MAVL/StickWRLD in Java3D (http://java.sun.com/products/java-media/3D/). The new implementation has three significant areas of improvement over the VRML version: enhanced real-time user interaction; clustering of residues by aggregate physicochemical properties and availability of pre-calculated StickWRLD WRLDs for all families in the PFAM database (3).

Our Java3D implementation leaves the computationally intensive calculation of the complete interpositional correlation matrix on the server, but moves the selection of features for display, to the client applet running on the user's computer. It is no-longer necessary to completely re-compute the correlation matrix to change display preferences. Therefore the user interface is no longer limited to a static graph with pre-applied statistical and display parameter choices. Instead, the various StickWRLD statistical parameters (Tr, the global over/underpopulation threshold; Pr, the per-edge overpopulation threshold; Nr, the per-edge underpopulation threshold; and alpha, the edge-significance threshold), and display parameters such as residue coloring and grouping can be adjusted by the user on-the-fly. This allows the parameters to be adjusted to suit the complexity of the alignment being visualized, and the user can now toggle between coarse visualization parameters to quickly explore and locate interesting features of an alignment, and fine parameters to investigate detailed aspects of the relationship.

To allow the system to more intuitively deal with the potentially weak relationships amongst poorly constrained sequences, clustering of residues by aggregate physicochemical properties has been implemented. This allows, for example, the conserved residue properties and interpositional requirements of distant homologs to be discovered, even in the absence of specifically conserved residues. Residues can currently be grouped by hydropathy, charge and volume, and the relationship between these aggregate properties at each position visualized, just as the specific sequence identities can be. Using this option in many protein families results in the discovery of correlated physical properties, even when the contributing residue identities fall below the significance threshold, or are obscured by other more dramatic individual-residue relationships.

Finally, we are interested in the distribution of the variety of interpositional relationships highlighted by MAVL/StickWRLD, amongst known sequence families. To facilitate access to this information, and to make it easier for researchers to examine popular sequence families, we have pre-computed StickWRLD WRLDs for the complete PFAM database of sequence families. The PFAM alignments, and the WRLDs generated from them, are available directly from the MAVL/StickWRLD Java interface. We currently house 7868 sequence families (the average number of sequences per family is 23 and average sequence length is 227) from PFAM, and we update the StickWRLD database regularly.

Some interesting statistics have come out of our examination of this data collection: The average number of possible correlations per family is approximately 1 000 000. On average, 5% of these display over/underpopulations >10% of expected population (Tr > 0.10). However, >60% of those with Tr > 0.10 have a significance better than 0.0001. This strongly suggests that many of the relationships highlighted by MAVL analysis are biologically relevant, rather than simple random outliers discovered by the massive number of correlations considered.

RESULTS AND DISCUSSION

The above improvements in the implementation result in fast, easy and accurate examinations of sequence families. Enhanced user interaction and availability of pre-calculated StickWRLD WRLDs for PFAM alignments provide fast and convenient analysis. Grouping of the residues based on physicochemical properties supplies more information on conserved patterns within families.

Applying MAVL/StickWRLD to a typical protein domain results in improved understanding of the interacting residues, and the physical properties required to maintain structure or function. Figure 1 shows StickWRLD representations and corresponding three-dimensional protein structures for the PFAM family of integrin alpha cytoplasmic regions (PDB 1m8o shown). This family contains the short cytoplasmic region of integrin alpha chains, and has small, strongly conserved α-helix, followed by a generally acidic region (4,5).

MAVL/StickWRLD analysis (Figure 1a) suggests that there are two alternative preferred sequences within the family {R9 P10 P11 Q12 E13} or {Y9 K10 M12}. That is, while the canonical structure contains an α-helix, terminating in R9 and broken by prolines at P10 and P11, there is an alternate sequence motif that contains neither P10 nor P11, but instead a preferred tyrosine, lysine, methionine triplet at {9, 10, 12}. A tutorial on reading the StickWRLD diagrams produced by the Java version is available from the MAVL/StickWRLD help-files page (http://www.microbial-pathogenesis.org/stickwrld/tutorial/sticktut2.html).

Furthermore, grouping of the residues by charge, to elucidate bulk-property-related propensities (Figure 1b, 1c) reveals more molecular preferences of the distinct sequence subfamilies. The turn following the α-helix of the cytoplasmic domain appears to be stabilized by either 5 residues following the pattern {9+, 10NP, 11NP, 12P, 13−} or 8 residues {8+, 9P, 10+, 11−, 12NP, 13NP, 14P, 15−} (+:positively charged, −:negatively charged, P:polar, NP:nonpolar) There is also a preference in the first pattern for a slightly polar residue at 2, and against a non-polar residue at this position. A high-resolution copy of Figure 1, as well as links to live versions of the Java StickWRLD diagram are available from the MAVL/StickWRLD help-files page (http://www.microbial-pathogenesis.org/stickwrld/tutorial/sticktut2.html).

Strongly conserved (consensus) residues of this small domain are colored and labeled in Figure 1d, and participate in the 10-residue N-terminal α-helix of the domain. However, consensus methods only identify the remainder of the domain as 'acidic', and do not capture the sequence or property requirements, or suggest their involvement in stabilizing the hairpin. The canonical (proline containing) motif is highlighted in Figure 1e. It is clear that the positively charged arginine and negatively charged glutamate are brought

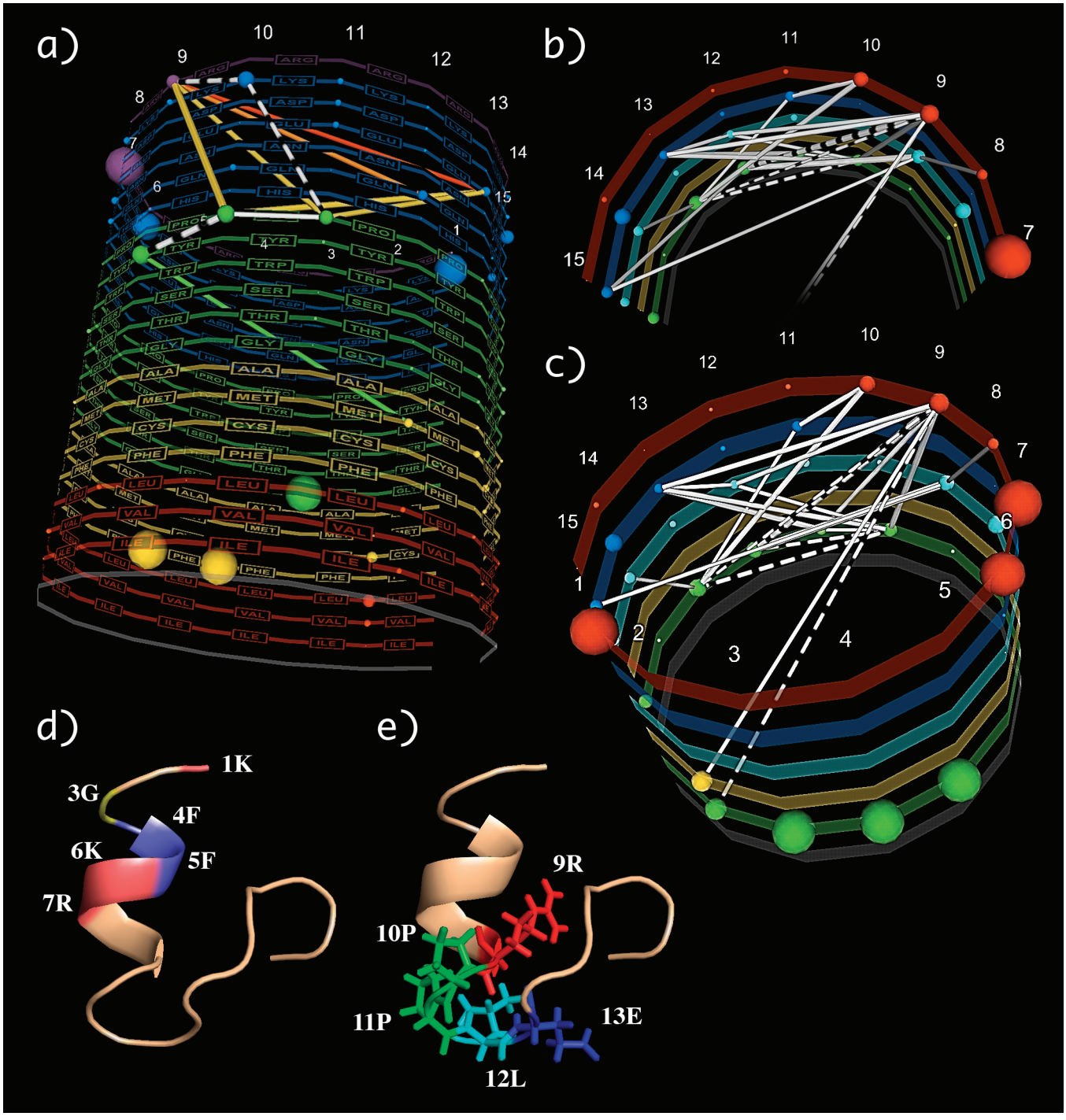

Figure 1. StickWRLD graph representations of the integrin alpha cytoplasmic domain family alignments, and a corresponding domain solution structure. **(a)** Shows interpositional correlations within the family in the form of a default StickWRLD graph; [**b** (detail) and **c** (overview)] show correlations obtained when residues are grouped by charge (red Positive, blue Negative, cyan Polar, tan Slightly-Polar, green Non-Polar, grey Gap); **(d)** illustrates the position of strongly conserved residues on the domain's structure **(e)** displays strongly correlating residues based on MAVL/StickWRLD analysis of the aligned members of the domain.

together by the hairpin, and can act in concert to stabilize the hairpin with the support of other polar and non-polar residues. While crystal or solution structures are not available for any non-proline-motif integrin alpha domains, we predict that the alternative positive residues at 8 and 10, and negative residues at 11 and 15 will be found to interact similarly to stabilize the domain structure.

CONCLUSIONS AND FUTURE WORK

MAVL/StickWRLD's new Java interface provides rapid exploration of alignment properties and insight into potential structural requirements that are embedded in the sequence identities. The analysis and visualization method has proven both intuitive and accurate for predicting likely

structural features in protein and nucleic acid families for which representative crystal or solution structures are known. We are developing additional utilities to assist the researcher by providing automatic suggested partitions of sequence families into subgroups, and are actively extending our pre-computed analysis database so that researchers may mine other interesting structural features from the PFAM families database.

ACKNOWLEDGEMENTS

Funding to pay the Open Access publication charges for this article was provided by the authors.

Conflict of interest statement. None declared.

REFERENCES

1. Ray,W. (2004) MAVL and StickWRLD: visually exploring relationships in nucleic acid sequence alignments. *Nucleic Acids Res*., **32**, W59–W63.
2. Ray,W. (2005) MAVL/StickWRLD for protein: visualizing protein sequence families to detect non-consensus features. *Nucleic Acids Res*., **33**, W315–W319.
3. Bateman,A., Coin,L., Durbin,R., Finn,R.D., Hollich,V., Griffiths-Jones,S., Khanna,A., Marshall,M., Moxon,S., Sonnhammer,E.L.L. *et al*. (2004) The Pfam Protein Families Database. *Nucleic Acids Res*., **32**, D138–D141.
4. Vinogradova,O., Velyvis,A., Velyviene,A., Hu,B., Haas,T.A., Plow,E.F. and Qin,J. (2002) A structural mechanism of integrin alfaIIb beta3 'inside-out' activation as regulated by its cytoplasmic face. *Cell*, **110**, 587–597.
5. Humphries,M.J., McEwan,P.A., Barton,S.J., Buckley,P.A., Bella,J. and Mould,A.P. (2003) Integrin Structure: heady advances in ligand binding, but activation still makes the knees wobble. *Trends Biochem. Sci*., **28**, 313–319.